A Fine-grained Perspective onto Object Interactions from First-person Views

Dima Damen[a]
Department of Computer Science, University of Bristol, Bristol, United Kingdom
Dima.Damen@bristol.ac.uk

Keywords: Egocentric Vision, First-Person Vision, Action Recognition, Fine-Grained Recognition, Object Interaction Recognition, Skill Determination, Action Completion, Action Anticipation, Wearable Cameras, First-Person Datasets, EPIC-Kitchens

Abstract: This extended abstract summarises the relevant works to the keynote lecture at VISAPP 2019. The talk discusses understanding object interactions from wearable cameras, focusing on fine-grained understanding of interactions on realistic unbalanced datasets recorded in-the-wild.

1 INTRODUCTION

Humans interact with tens of objects daily, at home (e.g. cooking/cleaning), during working (e.g. assembly/machinery) or leisure hours (e.g. playing/sports), individually or collaboratively. The field of research, within computer vision and machine learning, that focuses on the perception of object interactions from a wearable cameras is commonly referred to as 'first-person vision'. In this extended abstract, we cover novel research questions, particularly related to the newly released largest dataset in object interactions, recorded in people's native environments: EPIC-Kitchens.

2 Definitions

Object interactions could be perceived from different ordinal-person viewpoints - where 'ordinal' is used to generalise between *first-*, *second-* and *third-*person views. A view is referred to as a first-person view, if the interaction is captured by a wearable sensor, worn by the actor performing the interaction itself. Conversely, a second-person view is when the interaction is captured by a camera of a co-actor, or a recepient of the action. Finally, a third-person view, common in remote static cameras, is when the interaction is captured by an observer not relevant to the interaction or the actor during that interaction.

[a] https://orcid.org/0000-0001-8804-6238

3 Datasets and EPIC-Kitchens

For years, Computer Vision has focused on capturing videos from a third-person view, with the majority of action recognition datasets using a remote camera observing the action or interaction (Marszalek et al., 2009; Kuehne et al., 2011; Caba Heilbron et al., 2015; Carreira and Zisserman, 2017).

Increasingly, first-person vision datasets have been recorded, capturing full body motion such as sports (Kitani et al., 2011), social interactions (Alletto et al., 2015; Fathi et al., 2012a; Ryoo and Matthies, 2013) and object interactions (De La Torre et al., 2008; Fathi et al., 2012b; Pirsiavash and Ramanan, 2012; Damen et al., 2014; Georgia Tech, 2018; Sigurdsson et al., 2018).

In 2018, the largest dataset on wearable cameras was released through a collaboration led by the University of Bristol alongside the University of Catania and the University of Toronto - `http://epic-kitchens.github.io/`. EPIC-Kitchens (Damen et al., 2018) offers more than 11.5M frames, captured using a head-mounted camera in 32 different kitchens, with over 55 hours of natural interactions from cooking to washing the dishes (Fig 1).

4 Fine-Grained Object Interactions

Datasets, such as EPIC-Kitchens, can offer unique opportunities to studying previously unexplored prob-

Figure 1: Sample frames from EPIC-Kitchens

lems in fine-grained object interactions. A few of these opportunities are highlighted here.

- *Overlapping Object Interactions:* Defining the temporal extent of an action is fundamentally an ambiguous problem (Moltisanti et al., 2017; Sigurdsson et al., 2017). This is usually resolved through multi-labels, i.e. allowing a time-segment to belong to multiple classes of actions. However, actual understanding of interaction overlapping requires an space of action labels that captures dependencies (e.g. filling a kettle requires opening the tap). Models that capture and predict overlapping interactions are needed for a finer-understanding of object interactions.

- *Object Interaction Completion/Incompletion:* Beyond classification and localisation, action completion/incompletion is the problem of identifying whether the action's goal has been successfully achieved, or merely attempted. This is a novel fine-grained object interaction research question proposed in (Heidarivincheh et al., 2016). This work has been recently extended to locating the moment of completion (Heidarivincheh et al., 2018) - that is the moment in time beyond which the action's goal is believed to be completed by a human observer.

- *Skill Determination from Video:* Even when an interaction is successfully completed, further understanding of 'how well' the task was completed would offer knowledge beyond pure classification. In this leading work (Doughty et al., 2018a), a collection of video could be ordered by the skill exhibited in each video, through deep pairwise ranking. This method has been recently extended to include rank-aware attention (Doughty et al., 2018b) - that is a novel loss function capable of attending to parts of the video that exhibit higher skill as well as parts that demonstrate lower skill including mistakes or hesitation.

- *Anticipation and Forecasting:* Predicting upcoming interactions has recently gathered additional attention, triggered by the presence of first-person datasets (Furnari et al., 2018; Rhinehart and Kitani, 2017). Novel research on uncertainty in anticipating actions (Furnari et al., 2018), or relating forecasting to trajectory prediction (Rhinehart and Kitani, 2017) have recently been proposed.

- *Paired Interactions:* One leading work has attempted capturing both the action and its counter-action (or reaction), both from a wearable camera (Yonetani et al., 2016). This is a very exciting area of research, still under-explored.

5 Conclusion

Recent deep-learning research has only scratched the surface of potentials for finer-grained understanding of object interactions. As new hardware platforms for first-person vision emerge (Microsoft's Hololens, Magic Leap, Samsung Gear, $\cdots$), applications of fine-grained recognition will be endless.

REFERENCES

Alletto, S., Serra, G., Calderara, S., and Cucchiara, R. (2015). Understanding social relationships in egocentric vision. In *Pattern Recognition*.

Caba Heilbron, F., Escorcia, V., Ghanem, B., and Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., and Wray, M. (2018). Scaling egocentric vision: The EPIC-KITCHENS Dataset. In *ECCV*.

Damen, D., Leelasawassuk, T., Haines, O., Calway, A., and Mayol-Cuevas, W. (2014). You-do, I-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*.

De La Torre, F., Hodgins, J., Bargteil, A., Martin, X., Macey, J., Collado, A., and Beltran, P. (2008). Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. In *Robotics Institute*.

Doughty, H., Damen, D., and Mayol-Cuevas, W. (2018a). Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. In *CVPR*.

Doughty, H., Mayol-Cuevas, W., and Damen, D. (2018b). The Pros and Cons: Rank-aware temporal attention for skill determination in long videos. In *Arxiv*.

Fathi, A., Hodgins, J., and Rehg, J. (2012a). Social interactions: A first-person perspective. In *CVPR*.

Fathi, A., Li, Y., and Rehg, J. (2012b). Learning to recognize daily actions using gaze. In *ECCV*.

Furnari, F., Battiato, S., and Farinella, G. (2018). Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *ECCVW*.

Georgia Tech (2018). Extended GTEA Gaze+. http://webshare.ipat.gatech.edu/coc-rim-wall-lab/web/yli440/egtea_gp.

Heidarivincheh, F., Mirmehdi, M., and Damen, D. (2016). Beyond action recognition: Action completion in RGB-D data. In *BMVC*.

Heidarivincheh, F., Mirmehdi, M., and Damen, D. (2018). Action completion: A temporal model for moment detection. In *BMVC*.

Kitani, K. M., Okabe, T., Sato, Y., and Sugimoto, A. (2011). Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). HMDB: A large video database for human motion recognition. In *ICCV*.

Marszalek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In *CVPR*.

Moltisanti, D., Wray, M., Mayol-Cuevas, W., and Damen, D. (2017). Trespassing the boundaries: Labeling temporal bounds for object interactions in egocentric video. In *ICCV*.

Pirsiavash, H. and Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *CVPR*.

Rhinehart, N. and Kitani, K. M. (2017). First-person activity forecasting with online inverse reinforcement learning. In *ICCV*.

Ryoo, M. S. and Matthies, L. (2013). First-person activity recognition: What are they doing to me? In *CVPR*.

Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A., and Alahari, K. (2018). Charades-ego: A large-scale dataset of paired third and first person videos. In *ArXiv*.

Sigurdsson, G. A., Russakovsky, O., and Gupta, A. (2017). What actions are needed for understanding human actions in videos? In *ICCV*.

Yonetani, R., Kitani, K. M., and Sato, Y. (2016). Recognizing micro-actions and reactions from paired egocentric videos. In *CVPR*.